Репликация экспериментального поведения в игре доверия с использованием LLM

Пётр Паршаков, Юлия Найденова, Никита Маткин
НИУ ВШЭ - Пермь

Исследование выполнено за счет гранта Российского научного фонда № 25-18-00539, https://rscf.ru/project/25-18-00539/

Актуальность и проблема исследования

- LLM все чаще рассматриваются как синтетические респонденты и участники поведенческих экспериментов.
- Потенциальное преимущество: быстрое получение ответов при варьировании условий эксперимента.
- Методологическая проблема: правдоподобный ответ не обязательно означает воспроизведение поведения людей.
- Для экспериментальной экономики особенно важен вопрос внешней валидности таких симуляций.

Игра доверия как эмпирический тест

Механика игры

1. Первый игрок получает начальный ресурс.
2. Он выбирает сумму передачи второму игроку.
3. Переданная сумма утраивается.
4. Второй игрок решает, какую часть вернуть.

Что проверяется

Доверительное действие, ожидания относительно партнера и ответная взаимность второго игрока.

Почему это важно

LLM должны воспроизвести не одно число, а согласованную поведенческую структуру.

Berg et al. [1995]; Fehr [2009]; Alós-Ferrer, Farolfi [2019]

Литература, пробел и вклад

Игра доверия

Классическая линия: Berg et al. [1995], Ashraf et al. [2006], Bahry et al. [2003], Fehr [2009].

Игра позволяет отдельно смотреть доверие, ожидания и взаимность.

LLM как участники

LLM уже проверяют в опросах, сценариях и экономических играх: Aher et al. [2022], Filippas et al. [2023], Bisbee et al. [2024], Cui et al. [2024].

Близкие работы авторов

Alekseenko et al. [2025]: beauty contest и стратегическое поведение LLM.

Паршаков и соавт. [2026]: игра диктатора, LLM и люди.

Пробел: российский контекст в этой литературе представлен ограниченно; особенно мало понятно, воспроизводят ли LLM региональные различия внутри РФ.

Вклад работы: проверка LLM в игре доверия на опубликованных российских экспериментальных данных.

Исследовательский вопрос

Способны ли LLM воспроизводить поведение реальных участников в игре доверия в российском контексте?

- Как LLM воспроизводят действие первого игрока: сумму передачи?
- Как LLM воспроизводят ожидания возврата?
- Как LLM ведут себя в роли второго игрока?
- Меняются ли ответы при региональном контексте Татарстана и Сахи?

Эмпирическая стратегия

Ashraf et al. [2006]

Российская студенческая выборка, Москва.

Начальный капитал: 100 У.Е. = 1000 руб.

Решения: передача, ожидание возврата, возврат второго игрока.

Bahry, Whitt, Wilson [2003]

Татарстан и Республика Саха (Якутия).

Начальный капитал: 80 руб.

Решения: передача, ожидания, возврат и региональные различия.

Оба исследования позволяют проверять LLM на опубликованных российских экспериментальных данных.

Дизайн генерации ответов LLM

- 8 моделей: Claude Opus 4.6, Gemini 3.1 Pro, Grok 4.20, GPT-5.4, Kimi K2.5, GLM-5, YandexGPT 5.1 Pro, Gigachat 2 Max.
- Синтетические профили формировались на основе характеристик участников исходных экспериментов.
- Запрос включал профиль, роль участника, правила игры, допустимые решения и требуемый формат ответа.
- Ответы проверялись на допустимый диапазон и дискретный шаг.

Логика сопоставления: две статьи

Ashraf et al. [2006]

Слайды 9-11.

Российская студенческая выборка, Москва.

Проверяем три блока:
1. передача первого игрока;
2. ожидание возврата;
3. возврат второго игрока.

Bahry, Whitt, Wilson [2003]

Слайд 12.

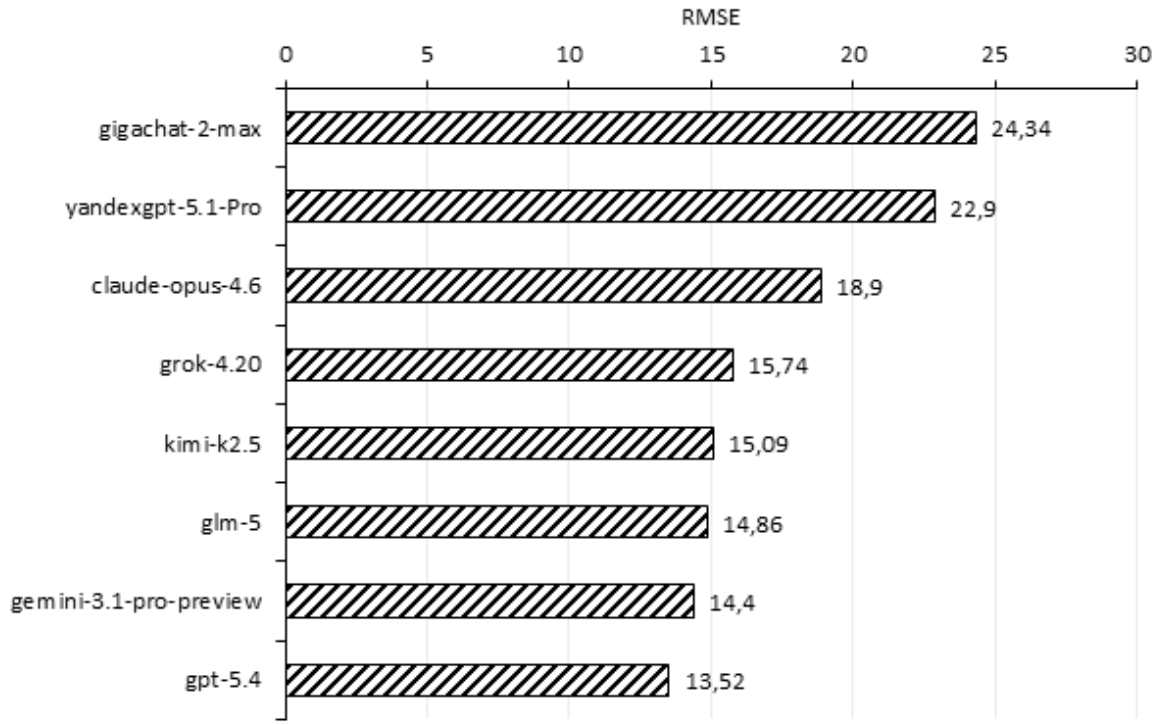
Татарстан и Республика Саха (Якутия).

Проверяем региональный контекст:
1. распределения передач;
2. отличие средних возвратов;
3. чувствительность к региону.

Идея: не смешивать источники — первые три результата относятся к Ashraf, региональный блок относится к Bahry.

Результат 1: действие первого игрока

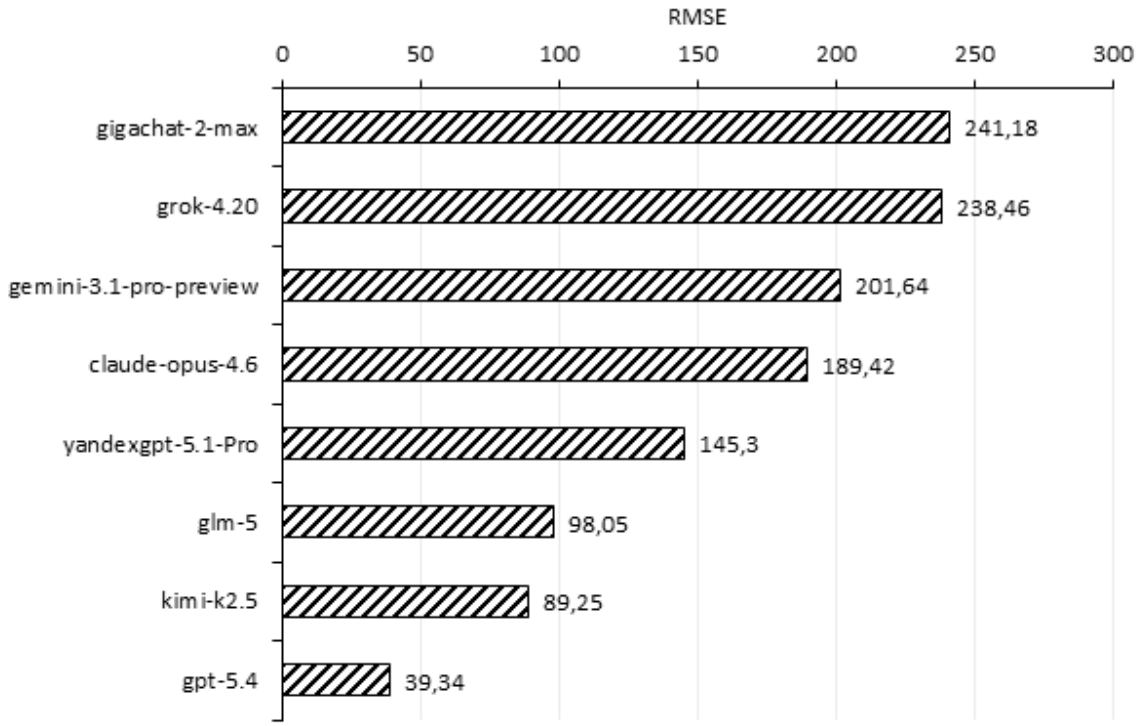
Ashraf et al. [2006]: российская студенческая выборка, Москва



- Столбец = RMSE: средняя ошибка распределения LLM относительно людей; 0 означает полное совпадение.
- Короткий столбец = ближе к эксперименту; длинный = сильнее отклонение. Сравнивать нужно внутри рисунка.
- Здесь диапазон 13,5-24,3: даже лучшая модель не совпадает с людьми, но ошибки различаются почти вдвое.

Результат 2: ожидания возврата

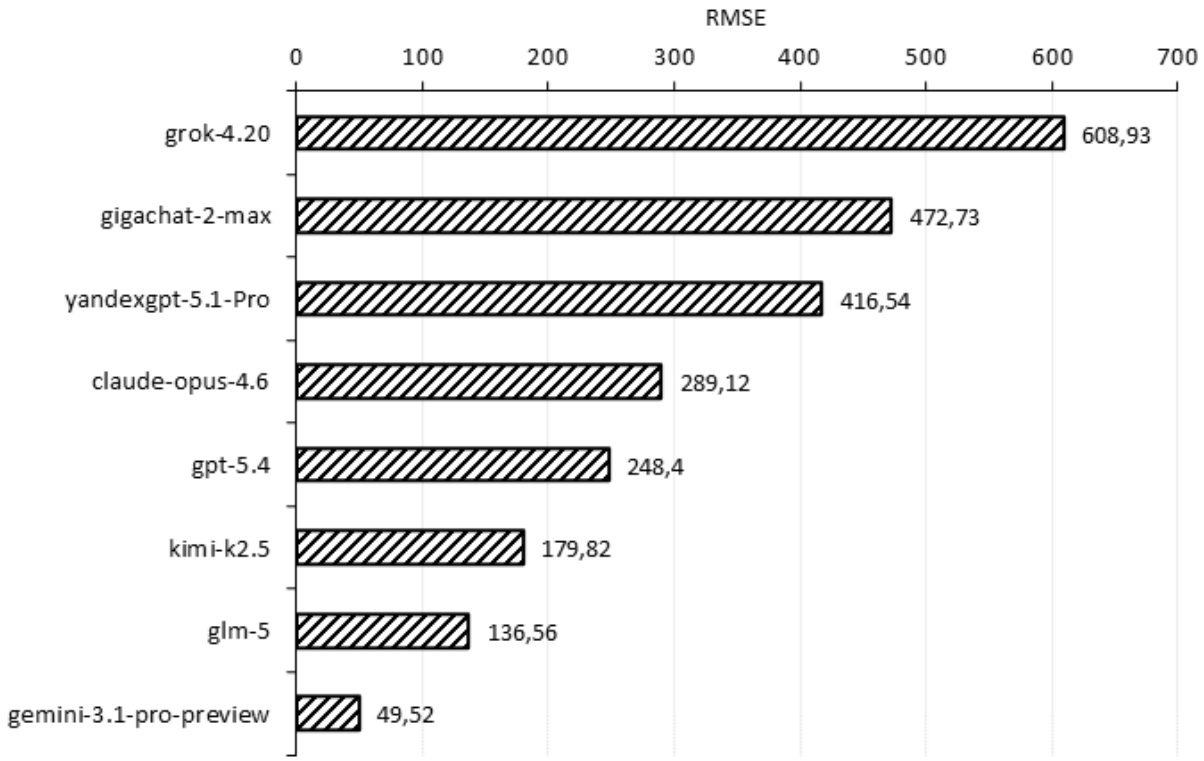
Ashraf et al. [2006]: российская студенческая выборка, Москва



- RMSE здесь считает ошибку в ожидаемом возврате; шкала другая, поэтому числа нельзя сравнивать с рис. 1 напрямую.
- 39 у GPT-5.4 = относительно близко; 200+ = крупное расхождение ожиданий с данными людей.
- Вывод: субъективные ожидания о партнере воспроизводятся хуже, чем сама передача.

Результат 3: взаимность второго игрока

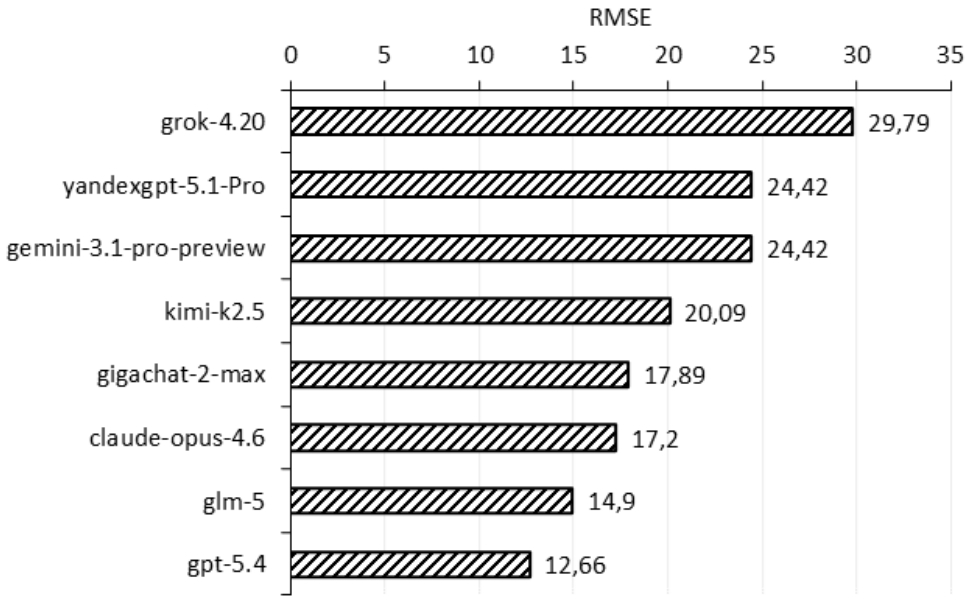
Ashraf et al. [2006]: российская студенческая выборка, Москва

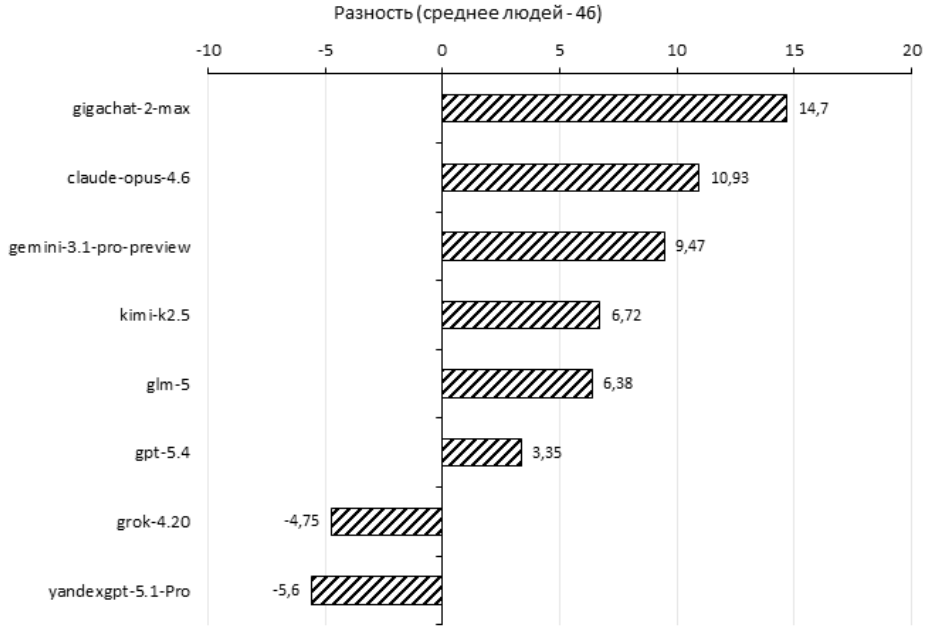


- RMSE измеряет ошибку суммы возврата второго игрока; 0 = совпадение со стратегией людей.
- 49,5 у Gemini = минимальное отклонение; 400-600+ = очень крупное расхождение по взаимности.
- Вывод: модели могут быть похожи в доверии, но сильно расходиться в ответной взаимности.

Результат 4: российский региональный контекст

Bahry, Whitt, Wilson [2003]: Татарстан и Республика Саха (Якутия)





Как читать: рис. 4 — меньше RMSE = ближе к региональным распределениям.
Рис. 5 — плюс/минус = выше/ниже среднего у людей; ближе к 0 = точнее.

Вывод: российский контекст не сводится к региону в промпте; качество зависит от модели и аспекта поведения.

Что следует из результатов

Где ближе к данным

В отдельных задачах модели дают ответы, похожие на экспериментальные данные людей, но это не устойчиво для всех моделей и ролей.

Где расходятся

Ожидания возврата, взаимность и региональные различия воспроизводятся нестабильно и зависят от модели.

Главный вывод

На российских данных LLM нельзя рассматривать как прямую замену участникам экспериментов без отдельной валидации.

- LLM полезны для пилотирования дизайна и предварительной проверки сценариев.
- В подтверждающих исследованиях нужны данные людей или калибровка на таких данных.
- Для нашей работы ключевой вклад — проверка поведения LLM в российском контексте игры доверия.